A discussion on learning and prior ignorance for sets of priors in the one-parameter exponential family

Alessio Benavoli and Marco Zaffalon
IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland
email: alessio@idsia.ch, zaffalon@idsia.ch

Abstract

For a conjugate likelihood-prior model in the one-parameter exponential family of distributions, we show that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors $\mathscr{M}$ that guarantees prior near-ignorance without producing vacuous inferences. This result is obtained following both a behavioural and a sensitivity analysis interpretation of prior near-ignorance. We also discuss the problem of the incompatibility of learning and prior near-ignorance for sets of priors in the one-parameter exponential family of distributions in the case of imperfect observations. In particular, we prove that learning and prior near-ignorance are compatible under an imperfect observation mechanism if and only if the support of the priors in $\mathscr{M}$ is the whole real axis.

Keywords. Prior near-ignorance, set of distributions, exponential family of distributions.

1 Introduction

This paper deals with the problem of modelling prior ignorance about statistical parameters through a set of prior distributions $\mathscr{M}$. There are two distinct approaches of this kind. The first approach, known as *Bayesian sensitivity analysis* [2], assumes that there is an ideal prior distribution π_0 which could, ideally, model prior uncertainty. It is assumed that we are unable to determine π_0 accurately because of limited time or resources. The criterion for including a particular prior distribution π in $\mathscr{M}$ is that π is a plausible candidate to be the ideal distribution π_0.

The second approach, known as the theory of coherent lower (and upper) previsions, was developed by Walley [11]. This approach revises Bayesian sensitivity analysis by directly emphasizing the upper and lower expectations (also called previsions) that are generated by $\mathscr{M}$. The upper and lower expectations of a bounded real-valued function (we call it a gamble) g on a possibility space, denoted by $\underline{E}(g)$ and $\overline{E}(g)$, are respectively the supremum and infimum of the expectations $E_P(g)$ over the probability measures P in $\mathscr{M}$ (if $\mathscr{M}$ is assumed to be closed and convex,[1] it is fully determined by all the upper and lower expectations). The upper and lower expectations have a behavioural interpretation (explained in Section 2), but, contrary to the sensitivity analysis approach, there is no special commitment to the individual probability distributions in $\mathscr{M}$. In choosing a set $\mathscr{M}$ to model prior near-ignorance, the main aim is to generate upper and lower expectations with the property that $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$ on a specific class of gambles of interest g. This means that the only available information about $E(g)$ is that it belongs to $[\inf g, \sup g]$, which is equivalent to state a condition of complete prior ignorance about the value of g.

Modeling a state of prior ignorance about the value w of a random variable W is not the only requirement for $\mathscr{M}$, it should also lead to non-vacuous posterior inferences. Posterior inferences are vacuous if the lower and upper expectations of all gambles of interest g coincide with the infimum and, respectively, the supremum of g. This means that our prior beliefs do not change with experience (i.e., there is no learning from data).

In [1], following an approach based on the behavioural interpretation, we have defined a set of *minimal properties* that a set $\mathscr{M}$ of distributions should satisfy to be a model of prior near-ignorance that does not lead to vacuous inferences. Furthermore, in the case that the likelihood model is in the one-parameter exponential family and $\mathscr{M}$ includes the corresponding conjugate exponential priors, we have also shown that the set of priors $\mathscr{M}$ satisfying the above properties can be uniquely obtained by letting the parameters of the conjugate exponential prior vary in suitable sets.

In this paper, after reviewing the main results of [1], we show that, for the one-parameter exponential family, similar conclusions about the parametrization of $\mathscr{M}$ (which guarantee prior near-ignorance and non-vacuous in-

[1] Closed and convex in the weak* topology, see [11, Sec. 3.6] for more details.

ferences) can be derived via a sensitivity analysis of the quantities of interest to the choice of the prior parameters.

We also deal with the problem of imperfect observations. In [8], it has been proven that the imprecise Beta model yields vacuous parametric inferences in the case the observation mechanism is imperfect. It is also shown that learning and prior near-ignorance are incompatible for the imprecise Beta model in the case of imperfect observations.[2] A question is if the impossibility to learn from imperfect observations under prior near-ignorance holds in general for any prior model based on sets of distributions. Here, considering conjugate likelihood-prior models in the one-parameter exponential family, we show that learning and prior near-ignorance are compatible under an imperfect observation mechanism if and only if the support of the priors in $\mathscr{M}$ is the whole real axis.

2 A Behavioural Interpretation of Prior Near-Ignorance

The aim of this section is to define which minimal properties the set of priors $\mathscr{M}$ should satisfy in the case where there is (almost) no prior information about $w \in \mathscr{W} \subseteq \mathbb{R}$. Before listing these properties, we discuss the behavioural interpretation of upper and lower expectations.

By regarding a gamble $g : \mathscr{W} \to \mathbb{R}$ as a random reward, which depends on the a priori unknown value of w, the expectation (also called prevision) of g w.r.t. w, i.e., $E(g)$, represents a subject's fair price for the function g. This means that he should be disposed to accept the uncertain rewards $g - E(g) + \varepsilon$ (i.e., to *buy* g at the price $E(g) - \varepsilon$) and $E(g) - g + \varepsilon$ (i.e., to *sell* g at the price $E(g) + \varepsilon$) for every $\varepsilon > 0$. More generally, the supremum acceptable buying price and the infimum acceptable selling prices for g need not coincide, meaning that there may be a range of prices $[a,b]$ for which our subject is neither disposed to buy nor to sell g at a price $k \in [a,b]$. His supremum acceptable buying price for g is then his lower expectation $\underline{E}(g)$, and it holds that the subject is disposed to accept the uncertain reward $g - \underline{E}(g) + \varepsilon$ for every $\varepsilon > 0$; and his infimum acceptable selling price for g is his upper prevision $\overline{E}(g)$, implying that he is disposed to accept the reward $\overline{E}(g) - g + \varepsilon$ for every $\varepsilon > 0$. A consequence of this interpretation is that $\underline{E}(g) = -\overline{E}(-g)$ for every gamble g.

Under this behavioural interpretation, a state of ignorance about a gamble g is modelled by setting $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$. This means that our subject is neither disposed to buy nor to sell g at any price $k \in [\inf g, \sup g]$. In other words, our subject is disposed to buy (sell) g only at a price strictly less (greater) than the minimum (maximum) reward that he would gain from g. This means that the available information on w does not allow our subject to set any meaningful buying or selling price for g, which is equivalent to stating that our subject is in a state of ignorance.

In [11], it is proven that a closed and convex set of probability distributions can be equivalently characterized by the lower (or upper) expectation functional that it generates as the lower (upper) envelope of the expectations obtained from the distributions in such a set. Vice versa, given a functional $\underline{E}(\cdot)$ that satisfies some regularity properties [11, Ch. 2], it is possible to define a family $\mathscr{M}$ of probability distributions that generates the lower expectation $\underline{E}(g)$ for any g. This establishes a one-to-one correspondence between closed convex sets of probability distributions and lower expectations.

In case the available prior information is scarce, it therefore seems more natural to define $\mathscr{M}$ according to the behavioural interpretation, i.e., in terms of the upper and lower expectations it generates [7]. For instance, in problems where there is (almost) no prior information one would expect the set $\mathscr{M}$ to be "large" in the sense that its generated upper and lower expectations are relatively far apart (vacuous or almost vacuous).

Modelling a state of prior ignorance about w is not the only requirement for $\mathscr{M}$, it must also produce non-vacuous posterior inferences (otherwise it is useless in practice). Hereafter, inspired by the work in [7], we define a set of minimal properties that $\mathscr{M}$ or, equivalently, the lower and upper expectations it generates, should satisfy to be a model of prior ignorance and produce consistent and meaningful posterior inferences. The first requirement for $\mathscr{M}$ is coherence.

(A.1) Coherence. Prior and posterior inferences based on $\mathscr{M}$ should be strongly coherent [11, Sec. 7.1.4(b)]. Under the behavioural interpretation, this means that we should not be able to raise the lower expectation (supremum acceptable buying price) of a given gamble g taking into account the acceptable transactions implicit in the other lower expectation models.

In practice, strong coherence imposes joint constraints on the prior, likelihood and posterior lower expectation models, in the sense that, when considered jointly, they should not imply inconsistent assessments. In [11, Sec. 7.8.1], it is proven that, in the case the prior and likelihood lower expectation models are obtained as lower envelopes of standard expectations w.r.t. sets of proper density functions and the posterior set of densities is obtained from these sets by element-wise application of Bayes' rule for density functions, then strong coherence of the

[2] Actually the results in [8] are more general and hold for a multivariate prior near-ignorant model defined on a compact set. However, since the present paper deals with the one-parameter exponential family, in the following we focus our attention on the restriction of [8] to the imprecise Beta model.

respective lower expectation models is satisfied.[3]

Besides coherence, other requirements for the set $\mathscr{M}$ are that it should represent the state of prior ignorance about w, but without producing vacuous posterior inferences. Thus, $\mathscr{M}$ should be large enough to model a state of prior ignorance w.r.t. a set of suitable gambles (i.e., a set of gambles of interest $\mathscr{G}_0$ w.r.t. which we assess our state of prior ignorance), but not too large to prevent learning from taking place. These two contrasting requirements are captured by the following two properties for $\mathscr{M}$.

(A.2) $\mathscr{G}_0$-prior ignorance. The prior upper and lower expectations of some suitable set of gambles $\mathscr{G}_0$ under $\mathscr{M}$ are vacuous, i.e., $\underline{E}[g] = \inf g(w)$ and $\overline{E}[g] = \sup g(w)$ for all $g \in \mathscr{G}_0$.

(A.3) $\mathscr{G}$-learning. For a chosen set of gambles $\mathscr{G} \supseteq \mathscr{G}_0$ and for each $g \in \mathscr{G}$ satisfying $\overline{E}[g] - \underline{E}[g] > 0$, there exists a finite $\delta > 0$ (possibly dependent on g) such that for each $n \geq \delta$ and non-empty sequence of observations $y^n = (y_1, \ldots, y_n)$, at least one of these two conditions is satisfied:

$$\underline{E}[g|y^n] \neq \underline{E}[g], \quad \overline{E}[g|y^n] \neq \overline{E}[g], \tag{1}$$

where $\underline{E}[\cdot|y^n]$ and $\overline{E}[\cdot|y^n]$ denote the posterior lower and upper expectations of g after having observed $y_1, \ldots, y_n$. Furthermore, for each $g \in \mathscr{G}_0$, (1) must hold for any $n > 0$.

Property (A.2) states that $\mathscr{M}$ should be vacuous a priori w.r.t. some set of gambles $\mathscr{G}_0$, i.e., the lower and upper expectations of $g \in \mathscr{G}_0$ respectively coincide with the infimum and the supremum of g. In case $\mathscr{M}$ includes all possible distributions then (A.2) holds for any function g. Here, conversely, we require that (A.2) is satisfied for some subset of gambles $\mathscr{G}_0$. The subset of gambles $\mathscr{G}_0$ used in (A.2) should include the gambles g w.r.t. which we state our condition of prior near-ignorance. Furthermore, the set $\mathscr{G}_0$ should be as large as possible to guarantee that also $\mathscr{M}$ is as large as possible, but no too large to be incompatible with the requirement (A.3) of learning. In fact, property (A.3) states that $\mathscr{M}$ should be non-vacuous a posteriori for any gamble $g \in \mathscr{G} \supseteq \mathscr{G}_0$, which is a condition for learning from the observations. The set of gambles $\mathscr{G}$ used in (A.3) should include the gambles g w.r.t. which we are interested in computing expectations (i.e., making inferences). The fact that $\mathscr{G}$ must include $\mathscr{G}_0$ is the only constraint on $\mathscr{G}$, meaning that (A.3) requires that $\mathscr{M}$ is not vacuous w.r.t. all these gambles for which the prior near-ignorance has been imposed. Moreover, for these gambles, it is required that (1) holds for any $n > 0$, i.e., after one observation the condition of prior-ignorance must already be left.

[3] This holds under standard assumptions about the existence of density functions and the applicability of Bayes' rule.

Since $\mathscr{M}$ is a model of prior near-ignorance, it is also desirable that the influence of $\mathscr{M}$ on the posterior inferences vanishes with increasing numbers of observations n. This is captured by the following property.

(A.4) Convergence. For each gamble $g \in \mathscr{G}$ and non-empty sequence of observations $y^n = (y_1, \ldots, y_n)$, the following conditions are satisfied for $n \to \infty$:

$$\begin{array}{rcl} \underline{E}[g|y^n] & \to & \underline{E}^*[g|y^n], \\ \overline{E}[g|y^n] & \to & \overline{E}^*[g|y^n], \end{array} \tag{2}$$

where $\underline{E}^*[g|y^n], \overline{E}^*[g|y^n]$ are the posterior lower and upper expectations obtained as lower envelopes of standard expectations w.r.t. the posterior densities derived, via Bayes' rule, from the likelihood model and the improper prior density $p(w) = 1$ for all $w \in \mathscr{W}$.

Property (A.4) states that, for $n \to \infty$, $\mathscr{M}$ should give the same lower and upper expectations of $g \in \mathscr{G}$ as those obtained from the improper prior density $p(w) = 1$. The fact that $\underline{E}^*[g|y^n] < \overline{E}^*[g|y^n]$ accounts for the general case in which the likelihood model is described by a set of likelihoods (for a single likelihood it would be $\underline{E}^*[g|y^n] = \overline{E}^*[g|y^n] = E^*[g|y^n]$). Although improper priors produce posteriors which are often incoherent with the likelihood model, (A.4) does not conflict with the requirement of coherence in (A.1). In fact (A.4) is a limiting property that holds only for $n \to \infty$ (furthermore, incoherence usually vanishes at the limit). In order to better understand properties (A.1)–(A.4), we show their instantiation for the case of the exponential family in Section 4. Before discussing these results, in the next section we introduce the exponential families of densities and review their main properties [4, Ch. 5].

3 Exponential Families

Consider a sampling model where i.i.d. samples of a random variable Z are taken from a sample space $\mathscr{Z}$.

Definition 1. A probability density $p(z|x)$, parametrized by $x \in \mathscr{X} \subseteq \mathbb{R}$, is said to belong to the one-parameter exponential family if it is of the form

$$p(z|x) = f(z)[g(x)]^{-1} \exp\left(c\phi(x)h(z)\right),\ z \in \mathscr{Z} \tag{3}$$

where, given f, h, ϕ and c, it results that $g(x) = \int_{z \in \mathscr{Z}} f(z) \exp\left(c\phi(x)h(z)\right) dz < \infty$. ■

Sometimes it is more convenient to rewrite (3) in a different form.

Definition 2. The probability density

$$p(y|w) = k(y) \exp(yw - b(w)),\ y \in \mathscr{Y}_m, \tag{4}$$

derived from (3) via the transformations $y = h(z)$, $\mathscr{Y}_m = h(\mathscr{Z})$, $w = c\phi(x)$, $b(w) = \ln(g(x))$ and $k(y) = f(z)$, is

called the canonical form of representation of the exponential family; w is called the natural (or canonical) parameter. ■

The canonical form has some useful properties. The mean and variance of Y are given by

$$E[Y|w] = \frac{db}{dw},\ E[(Y - E_Y[Y|w])^2|w] = \frac{d^2b}{dw^2}, \quad (5)$$

where it has been assumed that $\frac{d^2b}{dw^2}(w) > 0$; from (5) it follows that $\frac{db}{dw}(w) \in Int(\mathscr{Y})$ (i.e., interior of $\mathscr{Y}$) [5], where $\mathscr{Y} \subseteq \mathbb{R}$ is the smallest closed or semi-closed set that includes the sample mean of Y (if it exists, otherwise $\mathscr{Y} = Int(\mathscr{Y})$). Notice that the domain of the observations $\mathscr{Y}_m$ can be discrete or continuous, while $\mathscr{Y}$ is always continuous. In the case of n i.i.d. observations $y_i = h(z_i)$, it follows that

$$p(y^n|w) = \prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i)\exp(n(\hat{y}_n w - b(w))), \quad (6)$$

where $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the y_i which, together with n, is a sufficient statistic of y^n for inference about w under the i.i.d. assumption. Furthermore, by interpreting the density function in (6) as a likelihood function $L(w)$, with $y^n = (y_1, \ldots, y_n)$, we can define the corresponding conjugate prior.

Definition 3. A probability density $p(w|n_0, y_0)$, parametrized by $n_0 \in \mathbb{R}^+$ and $y_0 \in Int(\mathscr{Y})$, is said to be the canonical prior of (4) if

$$p(w|n_0, y_0) = k(n_0, y_0)\exp(n_0(y_0 w - b(w))), \quad (7)$$

where $w \in \mathscr{W}$, n_0 is the so-called number of pseudo-observations, y_0 is the so-called pseudo-observation and $k(n_0, y_0)$ is the normalization constant. ■

When $\mathscr{W} = \mathbb{R}$, $0 < n_0 < \infty$ and $y_0 \in Int(\mathscr{Y})$, (7) is a proper density [5]. Some examples of densities conjugate to a one-parameter exponential (canonical) family and defined in $\mathscr{W} = \mathbb{R}$ follow.

Gaussian with known variance: $y \in \mathscr{Y} = \mathbb{R}$, $x \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$,

$$p(y|x, \sigma^2) \propto \exp\left(-\tfrac{1}{2\sigma^2}(y-x)^2\right) \propto \exp\left(\tfrac{1}{\sigma^2}\left(yx - \tfrac{x^2}{2}\right)\right),$$

with $w = x$ and $b(w) = x^2/2$. The conjugate prior (7) transformed back to the original domain $\mathscr{X}$ is:

$$p(x|n_0, y_0) \propto \exp\left(-\tfrac{n_0}{2}(x - y_0)^2\right),$$

which is a Gaussian with mean y_0 and variance $1/n_0$.

Binomial-Beta: $x \in \mathscr{X} = (0,1)$, $y \in \{0,1\}$,

$$\begin{aligned} p(y|x) &\propto x^y(1-x)^{(1-y)} \\ &= (1-x)\exp\left(y\ln\left(\frac{x}{1-x}\right)\right) \\ &= \exp(yw - b(w)), \end{aligned}$$

$w = \ln(x/(1-x))$, $b(w) = -\ln(1-x) = \ln(1+\exp(w))$. Considering the change of variable $dx = \exp(w)/(1+\exp(w))^2 dw$, the conjugate prior (7) transformed back to the original domain $\mathscr{X}$ is:

$$p(x|n_0, y_0) \propto x^{n_0 y_0 - 1}(1-x)^{n_0(1-y_0)-1}$$

which is a Beta density with $n_0 = s > 0$ and $y_0 = t \in (0,1)$.

The pair likelihood and conjugate prior in the canonical exponential family satisfies a set of interesting properties, most of them are particularly useful to represent the nature of the Bayesian "learning" process. A list of such properties is given in the following lemmas, whose proof is omitted (see [4, Ch. 5]).

Lemma 1. *For a pair of likelihood and conjugate prior in the canonical exponential family, it holds that:*

(i) *the posterior density for w is:*

$$p(w|n_p, y_p) = k(n_p, y_p)\exp(n_p(y_p w - b(w))), \quad (8)$$

where $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$;

(ii) *the predictive density for future observations $(y_{n+1}, \ldots, y_{n+m})$ is*

$$p(y_{n+1}, \ldots, y_{n+m}|y_1, \ldots, y_n) = \prod_{j=1}^{m} k(y_{n+j}) \frac{k\left(n_0 + n, \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}\right)}{k\left(n_0 + n + m, \frac{n_0 y_0 + (n+m)\hat{y}_{n+m}}{n+m+n_0}\right)}. \quad (9)$$

■

Lemma 2. *Suppose that the canonical conjugate prior family is such that $p(w|n_0, y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$. Then the prior mean of the function $\frac{db}{dw}$ is $E\left[\frac{db}{dw}\middle| n_0, y_0\right] = y_0$ and the posterior mean is:*

$$E\left[\frac{db}{dw}\middle| n_p, y_p\right] = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}. \quad (10)$$

■

Notice that $p(w|n_0, y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$ holds for any canonical priors such that $\mathscr{W} = \mathbb{R}$, but in general it is not true for truncated priors, i.e., in the case $\mathscr{W} \subset \mathbb{R}$. This is one of the reasons why it has been assumed that $\mathscr{W} = \mathbb{R}$. In (5), it has been shown that $\frac{d}{dw}b(w)$ is the mean of Y. Hence, $\frac{d}{dw}b(w)$ is the quantity about which we will have prior beliefs before seeing the data y and posterior beliefs after observing the data. Hence, the results in Lemma 2 are particularly important, because they provide us with a closed formula for the prior and posterior mean of $\frac{d}{dw}b(w)$. For sampling models such that $\frac{d}{dw}b(w) = x$, i.e., linear exponential form (e.g., Gaussian, Beta and Gamma density), Lemma 2 gives thus a closed formula for the prior and posterior mean of x.

4 Sets of Conjugate Priors for Exponential Families

Consider the problem of statistical inference about the real-valued parameter w from noisy measurements $(y_1,\dots,y_n)$ and assume that the likelihood is completely described by the following probability density function (PDF) belonging to the exponential family:

$$\prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i)\exp(n(\hat{y}_n w - b(w))), \qquad (11)$$

where the parameters of the likelihood, i.e., sample mean $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $n \in \mathbb{R}^+$, are known (the likelihood can be modelled by a single PDF). By conjugacy and following a Bayesian approach, as prior for w we may consider the PDF $p(w|n_0,y_0)$ defined in (7) for a given value of the parameter y_0 and n_0. In the case there is not enough information about w to uniquely determine the values of the parameters y_0 and n_0, we can consider the family of priors $p(w|n_0,y_0)$ obtained by letting y_0 vary in $\mathscr{Y}' \subseteq Int(\mathscr{Y})$ and n_0 in some set $\mathscr{A}_{y_0} \subseteq \mathbb{R}^+$, which could depend on y_0. The question to be addressed is whether such family of priors satisfies the properties (A.1)–(A.4) discussed in Section 2. The answer to this question is given in the next theorem.

Theorem 1. *Consider as set of priors $\mathscr{M}$ the family of conjugate priors $p(w|n_0,y_0)$ with y_0 spanning the set $\mathscr{Y}' \subseteq Int(\mathscr{Y})$, n_0 spanning the set $\mathscr{A}_{y_0} \subseteq \mathbb{R}^+$ (with $\mathscr{A}_{y_0}$ possibly dependent on y_0), under the assumptions: $\mathscr{Y}$ convex and $\mathscr{W} = \mathbb{R}$. If and only if the following conditions hold:*

(a) For each $y_0 \in \mathscr{Y}'$ and $n_0 \in \mathscr{A}_{y_0}$, it holds that $p(w|n_0,y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$;

(b) $\mathscr{Y}' = Int(\mathscr{Y})$;

(c) $\mathscr{A}_{y_0}$ satisfies the following constraints: $0 < \inf\mathscr{A}_{y_0}$, $\sup\mathscr{A}_{y_0} \leq \min(\overline{n}_0, \frac{c}{|y_0|})$ for each $y_0 \in Int(\mathscr{Y})$ and given parameters $\overline{n}_0, c > 0$;

then, given the parameters $\overline{n}_0$ and c, $\mathscr{M}$ is the largest set which satisfies properties (A.1)–(A.4), with $\mathscr{G}_0 = \{\frac{db}{dw}\}$ and $\mathscr{G}$ including sufficiently smooth gambles.[4] ■

The proof of the theorem can be found in [1, Sec. 4]. Hereafter, we illustrate the intuition behind the theorem. We distinguish three cases $\mathscr{Y} = \mathbb{R}$, $\mathscr{Y} = [a,\infty)$ (or $\mathscr{Y} = (-\infty,a]$) with $a \in \mathbb{R}$, and $\mathscr{Y} \subset \mathbb{R}$ bounded. In the last two cases w.l.o.g. it can be assumed that $\mathscr{Y} = [0,\infty)$ (or $\mathscr{Y} = (-\infty,0]$) and, respectively, $\mathscr{Y} = [0,1]$ (by shifting and scaling $\mathscr{Y}$); since $\mathscr{Y}$ has been assumed to be convex, these three cases account for all the possibilities.

[4] With sufficiently smooth gambles, we mean integrable w.r.t. the exponential family density functions with support in $\mathscr{W}$ and continuous on a neighborhood of the point where the posterior relative to the improper prior $p(w) = 1$ concentrates for $n \to \infty$.

Consider the case in which the observations belong to $\mathbb{R}$ and the likelihood is a Gaussian density with known variance, so that $\mathscr{Y} = (-\infty,+\infty)$. The conjugate model under considerations is thus a Gaussian-Gaussian model. In this case, the set of priors $\mathscr{M}$ is equal to:

$$\Big\{\mathscr{N}\left(w;y_0,\sigma_0^2\right) :\ y_0 \in (-\infty,+\infty),\ \max(1/\overline{n}_0, |y_0|/c) < \sigma_0^2 < \infty\Big\}, \qquad (12)$$

where y_0 is the prior mean and $\sigma_0^2 = 1/n_0$ the prior variance. Hence, $\mathscr{M}$ includes all the Gaussian densities with mean free to vary in $\mathbb{R}$ and variance lower bounded by $1/\overline{n}_0$ but linearly increasing with $|y_0|$. Notice, in fact, that if $|y_0| > c/\overline{n}_0$, then $\sigma_0^2 \geq |y_0|/c$. Hence, considering the likelihood $\mathscr{N}(y_i;w,\sigma^2)$ for $i = 1,\dots,n$, the corresponding set of posteriors is equal to:

$$\Big\{\mathscr{N}\left(w;y_p,\sigma_p^2\right) :\ y_p = \sigma_p^2\left(\frac{y_0}{\sigma_0^2} + \frac{n\hat{y}_n}{\sigma^2}\right),\ \sigma_p^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1},\ y_0 \in (-\infty,+\infty),\ \max(1/\overline{n}_0, |y_0|/c) < \sigma_0^2 < \infty\Big\}, \qquad (13)$$

where y_p is the posterior mean. Since $y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0)$ then, fixed $n_0 = 1/\sigma_0^2$, for $|y_0| \to \infty$ it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n+n_0) = |y_0| \to \infty$. Similarly, fixed y_0, for $n_0 \to \infty$ it follows that $|y_p| = |y_0|$. In other words, $n_0|y_0| = \infty$ implies a vacuous posterior mean and, thus, no learning and no convergence. Theorem 1 states that a necessary and sufficient condition to guarantee near-ignorance without preventing learning and convergence to take place is by imposing the constraint:

$$|n_0 y_0| < c < \infty,$$

which means that n_0 must in general depend on y_0. In this case in fact for $|y_0| \to \infty$, it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n+n_0) < \infty$. That is, the contribution of y_0 to y_p must decrease as $|y_0| \to \infty$, otherwise the observations do not contribute to y_p (learning cannot take place). This is essentially the meaning of the constraint $|y_0|/c < \sigma_0^2$ in (13), i.e., the variance of the Gaussians in $\mathscr{M}$ must be greater than $|y_0|/c$. Furthermore, $n_0 < \infty$ or, equivalently, the variance must also be greater than zero otherwise the Gaussian density would coincide with a Dirac delta; this is the reason of the constraint $\sigma_0^2 > 1/\overline{n}_0 > 0$. Under these constraints, it can be verified that y_p satisfies:

$$\min\left(\frac{-c+n\hat{y}_n}{n+\overline{n}_0}, \frac{-c+n\hat{y}_n}{n}\right) \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \max\left(\frac{c+n\hat{y}_n}{n+\overline{n}_0}, \frac{c+n\hat{y}_n}{n}\right), \qquad (14)$$

and converges to $\hat{y}_n$ (maximum likelihood estimate) for $n \to \infty$ (convergence property (A.4)). Observe that, for

$\overline{n}_0$ suitably small, the set of priors $\mathscr{M}$ reduces to the family of Gaussian priors with infinite variance discussed in [7, Section 3.3] and the bounds in (14) become approximately equal to:

$$\frac{-c+n\hat{y}_n}{n} \leq \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \frac{c+n\hat{y}_n}{n}. \tag{15}$$

The main difference is that the family of priors defined in Theorem 1 has been proved to be strongly coherent, while no proof of coherence is given for the model in [7, Section 3.3]; the coherence of this model is still an open problem.

Consider now the case in which the observations are counts, i.e., the likelihood is a Poisson distribution, $\mathscr{Y}_m = \mathbb{N}$ and $\mathscr{Y} = [0,\infty)$. The conjugate model under consideration is now a Poisson-Gamma model. The set of priors $\mathscr{M}$ transformed back to the original parameter space $\mathscr{X}$ reduces to a set of Gamma densities:

$$\mathscr{M} = \Big\{ g(x|\alpha,\beta) : \ 0 < \alpha = n_0 y_0 \leq c, \\ 0 < \beta = n_0 \leq \min(\overline{n}_0, c/|y_0|) \Big\}, \tag{16}$$

where $x, y_0 \in (0,+\infty)$ and $g(x|\alpha,\beta) \propto x^{\alpha-1}\exp(-\beta x)$ is the Gamma density with parameters α and β. The set of posteriors resulting from (16) is:

$$\mathscr{M}_p = \Big\{ g(x|\alpha,\beta) : \ \alpha = n_0 y_0 + n\hat{y}_n, \ \beta = n+n_0, \\ y_0 \in (0,+\infty), \ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|) \Big\} \tag{17}$$

and the posterior mean is equal to $y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0)$. Notice again that, because of the constraint $n_0 \leq \min(\overline{n}_0, c/|y_0|)$ it results that y_p is always finite, satisfies[5]

$$\frac{n\hat{y}_n}{n+\overline{n}_0} \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \frac{c+n\hat{y}_n}{n},$$

and converges to $\hat{y}_n$ (maximum likelihood estimate) for $n \to \infty$.

Consider the case in which the observations are binary, i.e., the likelihood is a binomial distribution $\mathscr{Y}_m = \{0,1\}$ and $\mathscr{Y} = [0,1]$. The conjugate model under considerations is thus a Binomial-Beta model. It can be easily verified that in this case the set of priors $\mathscr{M}$ transformed back to the original parameter space $\mathscr{X}$ reduces to the general Imprecise Beta Model (IBM) discussed in [11, Section 5.4.3]:

$$\mathscr{M} = \Big\{ B(x; st, s(1-t)) : t \in (0,1), 0 < s < \overline{n}_0 \Big\}, \tag{18}$$

where $x \in (0,1)$, $y_0 = t$, $n_0 = s$ and $B(x;\alpha,\beta)$ is the Beta density with parameters α and β. In this case, it follows from Theorem 1 that $y_0 \in (0,1)$ and $0 < n_0 \leq \min(\overline{n}_0, c)$. Hence, if $\overline{n}_0 < c$ the set of priors in Theorem 1 reduces to (18). In this case, near-ignorance and learning/convergence are compatible even if n_0 does not depend on y_0. In fact, being $|y_0| < 1 < \infty$, the product $n_0 y_0$ is always bounded provided that $n_0 < \overline{n}_0 < \infty$.[6] Finally notice that in the special case $s = \overline{n}_0$, we obtain the IBM discussed in [11, Section 5.3.1] and [3].

Observe that the family of priors $\mathscr{M}$ in Theorem 1 is completely determined by the two parameters $c > 0$ and $\overline{n}_0 > 0$. The larger these parameters are the larger the family of priors $\mathscr{M}$ is and, thus, the more conservative are the posterior inferences. The choice of these parameters is discussed in [1, Sec. 5].

It is also interesting to compare the set of priors $\mathscr{M}$ in Theorem 1 with another model for near-ignorance, the Bounded Derivative Model (BDM) [12]. In the BDM, $\mathscr{M}_{BDM}$ includes all continuous proper probability density functions for which the derivative of the log-density is bounded by a positive constant. It can be verified that BDM satisfies all the properties (A1)–(A4), with $\mathscr{G}_0$ and $\mathscr{G}$ defined as in Theorem 1. BDM is a non-parametric model and, in this sense, is more general than the model resulting from Theorem 1 that is restricted to the one-parameter exponential family only. A drawback of this generality is that inferences with BDM can in general be difficult to compute [12, Sec. 6], while this is often not the case for the model resulting from Theorem 1 because of conjugacy.

Conversely, a model for statistical inferences based on a set of densities belonging to the exponential family is presented in [9, Ch.4], [10]. The main difference w.r.t. the present work is that the model in [10] is not a model of prior near-ignorance, as pointed out by the authors, i.e., the set $\mathscr{Y}'$ in Theorem 1 is chosen in [10] to reflect the prior information on y_0 and, thus, the posterior inferences depend on this information. Since no constraint between n_0 and y_0 is assumed, the model in [10] can also violate (A.3)–(A.4) in the case $\mathscr{Y}' = Int(\mathscr{Y})$, and hence it can produce vacuous inferences.

5 A Sensitivity Analysis Interpretation of Prior Near-Ignorance

In Section 2, we have considered an interpretation of prior near-ignorance in terms of lower and upper expectations, i.e., behavioural dispositions to buy and sell gambles. In particular, with the properties (A1)–(A4), we have given general conditions for coherence, prior near-ignorance, learning and convergence, which hold for any set of distributions $\mathscr{M}$. Then, in Section 4, we have specialized these

[5] Since $\hat{y}_n \geq 0$, it results that $\frac{c+n\hat{y}_n}{n} \geq \frac{c+n\hat{y}_n}{n+\overline{n}_0}$ and, thus, $\frac{c+n\hat{y}_n}{n}$ is a right bound for y_p.

[6] In [13] the authors propose a functional relationship between n_0 and y_0 in the exponential families with a different aim w.r.t. that of the present paper; that is highlighting prior-data conflict in the case of inference drawn from a set of informative priors, i.e., near-ignorance is not satisfied. In this case, n_0 may depend on y_0 also in the IBM.

conditions to the case in which $\mathscr{M}$ includes densities belonging to the one-parameter exponential family and, for this set of densities, we have shown that (A1)–(A4) are equivalent to a special choice of the domains for the parameters of the exponential priors.

An alternative approach is to start directly from the set of priors $\mathscr{M}$ in the one-parameter exponential family and then to perform a sensitivity analysis of the quantities of interest (posterior inferences) to the choice of the prior parameters. This is typically done by deriving the quantities of interest w.r.t. the parameters of the conjugate priors, and looking for a set of parameters that sharply changes the inferences.

In this respect, consider a function $g(\frac{db}{dw})$[7] and its Taylor series expansion around the posterior parameter y_p, i.e.:

$$g\left(\frac{db}{dw}\right) = g(y_p) + \left(\frac{db}{dw} - y_p\right) g'(y_p) + \frac{1}{2}\left(\frac{db}{dw} - y_p\right)^2 g''(y_p) + \dots \qquad (19)$$

where $g'(y_p) = \frac{dg}{d(\frac{db}{dw})}|_{y_p}$ and so on for higher order derivatives. In statistical inference, we are interested in computing the expectation of g or, equivalently, of (19) w.r.t. the posterior density $k(n_p, y_p)\exp(n_p(y_p w - b(w)))$, i.e.:

$$\begin{aligned} E[g|y^n] &= \int g\left(\frac{db}{dw}\right) k(n_p, y_p)\exp(n_p(y_p w - b(w)))dw \\ &= g(y_p) + \frac{1}{2}g''(y_p)E\left[\left(\frac{db}{dw} - y_p\right)^2 \middle| y^n\right] \\ &\quad + \frac{1}{3!}g'''(y_p)E\left[\left(\frac{db}{dw} - y_p\right)^3 \middle| y^n\right] + \dots \end{aligned} \qquad (20)$$

where, for short notation, $\{y_1, \dots, y_n\} = y^n$ has been introduced. The posterior expectation $E[g|y^n]$ depends on $y_p = (n_0 y_0 + n\hat{y}_n)/(n + n_0)$ which, in turn, depends on the prior parameters n_0 and y_0. The sensitivity of $E[g|y^n]$ to the prior parameters can be obtained by differentiating $E[g|y^n]$ w.r.t. n_0 and y_0. However, since the value of n_0 may depend on the value of y_0 and vice versa, it is more interesting to compute the sensitivity of $E[g|y^n]$ to variations of $n_0 y_0$. Define $n_0 y_0 = r$ and $n_0 = n_0(r)$, then

$$\frac{dy_p}{dr} = \frac{n + n_0 - (r + n\hat{y}_n)\frac{dn_0}{dr}}{(n + n_0)^2}. \qquad (21)$$

where n_0 depends on r. Thus, it follows that $\dfrac{dE[g|y^n]}{dr}$ is equal to

$$\begin{aligned} &\frac{dy_p}{dr}\frac{dg(y_p)}{dy_p} + \frac{1}{2}\frac{dy_p}{dr}\frac{dg''(y_p)}{dy_p}E\left[\left(\tfrac{db}{dw} - y_p\right)^2 \middle| y^n\right] \\ &+ \frac{1}{2}\frac{dy_p}{dr}g''(y_p)\frac{dE\left[\left(\frac{db}{dw} - y_p\right)^2 \middle| y^n\right]}{dy_p} + \dots \end{aligned} \qquad (22)$$

From the relationship between a derivative and its difference quotient, one gets

$$|E_{r+\Delta}[g|y^n] - E_r[g|y^n]| \leq \left|\frac{dE[g|y^n]}{dr}\right| |\Delta| \qquad (23)$$

where $E_{r+\Delta}[g|y^n]$ is the expected value of g computed at $n_0 y_0 = r + \Delta$, $E_r[g|y^n]$ is the expected value of g computed at $n_0 y_0 = r$ and Δ is a scalar such that $r + \Delta \in [\min n_0 y_0, \max n_0 y_0]$.

Theorem 2. *There exists a finite $\delta > 0$ (possibly dependent on g) such that, for each $n \geq \delta$ and non-empty set of observations $y_1, \dots, y_n$, the difference $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]|$ is bounded and converges to zero for $n \to \infty$, if $\max|n_0 y_0| < \infty$ and $n_0 < \infty$.* ■

Proof: *If $\max|n_0 y_0| < \infty$, then it is also true that $\max|\Delta| = |\max n_0 y_0 - \min n_0 y_0| < \infty$. With $\max|\Delta|$ being bounded, a condition for $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]|$ to be bounded is that $|dE[g|y^n]/dr| < \infty$. Thus, also being $n_0 < \infty$, for $n \to \infty$ it follows that $y_p \to \hat{y}_n$, $n_p \to n$ and the posterior density $p(w|n_p, y_p)$ becomes a Dirac delta in $\hat{y}_n$. Then it results that $\lim E\left[\left(\frac{db}{dw} - y_p\right)^m \middle| y^n\right] = 0$ for any $m = 1, 2, \dots$ and $\lim dy_p/dr = 0$ (since $y_p = \hat{y}_n$, the derivative of y_p w.r.t. r is null). Thus, $|dE[g|y^n]/dr|$ converges to zero for $n \to \infty$. Furthermore, because $p(w|n_p, y_p)$ is always a well-defined PDF if $|n_0 y_0| < \infty$ and $n_0 < \infty$, by continuity arguments we can also conclude that there exists a finite $\delta > 0$ such that $|dE[g|y^n]/dr|$ is bounded for any $n > \delta$.* ■

Thus, we have again proven that $\max|n_0 y_0| < c$ and $n_0 \leq \overline{n}_0 < \infty$ are sufficient conditions for learning and convergence,[8] but now following an approach based on sensitivity analysis. Consider the case $g\left(\frac{db}{dw}\right) = \frac{db}{dw}$, assume that $p(w|n_p, y_p)$ is a Beta density and $n_0 = s > 0$. Then, from (21)–(22) it follows that $dy_p/dr = dE[g|y^n]/dr = 1/(n + s)$ (because $n_0 = s$ is constant). Since $y_0 \in (0, 1)$, then $0 < n_0 y_0 = r < s$ and, thus, $\max|\Delta| = s$, we conclude that $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]| \leq \frac{s}{n+s}$, which is exactly the imprecision (i.e., the difference between the upper and lower mean) of the IBM.

Consider the Gaussian case and assume $n_0 \approx 0$. For $g\left(\frac{db}{dw}\right) = \frac{db}{dw}$ it results that $dy_p/dr = dE[g|y^n]/dr = 1/n$. In this case the boundedness of $\max|\Delta|$ is ensured if $|n_0 y_0| \leq c < \infty$, which implies $\max|\Delta| = 2c$. Therefore,

[7] To simplify the derivations, we have assumed that g is an analytic function. Although not general, this holds for many gambles g.

[8] Theorem 1 is more general than Theorem 2, since it holds for more general functions g. Furthermore, the conditions derived there are not only sufficient but also necessary for (A.1)–(A.4).

(23) becomes $\max_{r,\Delta}|E_{r+\Delta}[g|y^n]-E_r[g|y^n]|\leq\frac{2c}{n}$, which is the imprecision of (15). Therefore, we have arrived at similar conclusions of those in Theorem 1 but via a sensitivity analysis. This approach allows to give another interpretation of the imprecision, e.g., $s/(n+s)$ and $2c/n$, in terms of the maximum value of the product $|dE[g|y^n]/dr||\Delta|$.

6 Imperfect observations

In real world applications, there is always a probability of making mistakes during the observation process. Often, if this probability is small, one assumes that the data are perfectly observable in order to use a simple likelihood model (e.g., a density belonging to the exponential family); doing so, one implicitly assumes that there is a sort of continuity between models with perfectly observable data and models with small probability of errors in the observations. In other words, one expects that a small error in the modelling of the observation mechanism leads to a small error in the inference. However, as observed in [8], this may be not true for inferences derived from a prior near-ignorance model based on set of distributions. To better understand this aspect, we introduce the imperfect observation mechanism described in [8]. An imperfect observation mechanism can be modelled as a two step process: (i) ideal observations $y'_1,\ldots,y'_n$ are generated according to the likelihood $L(y'_1,\ldots,y'_n|w)$; (ii) $y'_1,\ldots,y'_n$ are perturbed based on a distribution $p(y_1,\ldots,y_n|y'_1,\ldots,y'_n)$ and imperfect observations $y_1,\ldots,y_n$ are produced. Hence, the likelihood of imperfect observations can be modelled as:

$$p(y^n|w)=\int_{\mathscr{Y}_m^n}p(y^n|y'^n)L(y'^n|w)\,dy'^n, \tag{24}$$

where, for the sake of space, the notation $y^n=(y_1,\ldots,y_n)\in\mathscr{Y}_m^n$ and $y'^n=(y'_1,\ldots,y'_n)\in\mathscr{Y}_m^n$ has been introduced; $p(y^n|y'^n)=\prod_{i=1}^n p(y_i|y'_i)$ is any PDF such that $p(y_i|y'_i)>0$ for all $y_i,y'_i\in\mathscr{Y}_m$; $L(y'^n|w)=\prod_{i=1}^n L(y'_i|w)$ is the likelihood corresponding to the ideal unknown observations y'_i (we assume that it belongs to one-parameter canonical exponential family of distributions). Since the observations can also be discrete, $p(y^n|y'^n)$ and $L(y'^n|w)$ can also be probability mass functions and the integral in (24) becomes a sum. For the sake of notation, we use the integral notation for both continuous and discrete case, but in the latter case (24) becomes:

$$p(y^n|w)=\sum_{y'\in\mathscr{Y}_m^n}p(y^n|y'^n)L(y'^n|w).$$

Assume we have no prior information about w and we use the model in Theorem 1 to represent our state of ignorance. Since $p(y_1,\ldots,y_n|w)$ might not belong to the exponential family of distributions, a question to be addressed is if properties (A3)–(A4) continue to hold also in this case. The answer is in general negative as shown in [8]. In fact, assuming the imperfect observation mechanism (24), the authors prove that, for the Imprecise Beta model (as discussed in the Introduction, the results in [8] are more general), property (A.3) does not hold (no learning from data takes place) and, consequently also (A.4) does not hold (no convergence). In this case, the only way to satisfy (A.3)–(A.4) is to not allow $y_0\to 0,1$; this means that y_0 must vary in $[\varepsilon,1-\varepsilon]$ with $0<\varepsilon<0.5$. That is, (A.3)–(A.4) can be satisfied if and only if (A.2) (prior near-ignorance) does not hold [8]. A similar conclusion is derived in [6] using more general arguments. This has an important consequence, namely that in this case, the amount of imperfection introduced by $p(y^n|y'^n)$ (as long as it is positive) does not matter, we cannot be ignorant a priori without also being vacuous a posteriori.

A further question to be addressed is if this is true for any conjugate model (e.g., Gaussian-Gaussian, Poisson-Gamma etc.), whose likelihood is perturbed as described in (24). In order to prove that, we will use the following results.

Lemma 3. *Consider the prior* $p(w|n_0,y_0)=k(n_0,y_0)\exp(n_0(y_0w-b(w)))$. *For* $y_0\to\sup\mathscr{Y}$ *or* $y_0\to\inf\mathscr{Y}$ *and* $n_0<\infty$, *it holds that* $k(n_0,y_0)\to 0$ *and* $\exp(n_0(y_0w-b(w)))$ *concentrates on the value* w^* *such that* $db(w)/dw|_{w=w^*}=y_0$. ■

This can be proven by using the same arguments in the proof of [1, Cor. 1] (notice that w^* is a maximum of $p(w|n_0,y_0)$).

Lemma 4. *Consider the observational mechanism (24) and assume that:* $p(y^n|y'^n)>0$ *for each* $y^n,y'^n\in\mathscr{Y}_m^n$, $L(y'^n|w)$ *belongs to the exponential family of distributions and* $\mathscr{W}=\mathbb{R}$. *Define* $Lg_n(w)=\ln p(w|n,y^n,y_0,n_0)=\ln(p(y^n|w)p(w|n_0,y_0)/p(y^n))$ *and assume that for any well-defined prior* $p(w|n_0,y_0)$, *with* $0<n_0<\infty$ *and* $y_0\in Int(\mathscr{Y})$, *and for every n there is a strict local maximum* m_n *of* $p(w|n,y^n,y_0,n_0)$ *satisfying:*

$$\frac{dLg_n}{dw}(m_n)=0,\ \sigma_n^2=-\left(\frac{d^2Lg_n}{dw^2}(m_n)\right)^{-1}>0 \tag{25}$$

and that m_n *converges when* $n\to\infty$. *Define* $B_\rho(w^*)=\{w:|w-w^*|<\rho\}$ *and assume also that the posterior satisfies:*

(c1) $\sigma_n^2\to 0$ *for* $n\to\infty$.

(c2) *For any* $\varepsilon>0$ *there exists* $\delta>0$ *and* $\rho>0$ *such that, for any* $n>\delta$ *and* $w\in B_\rho(m_n)$, *it holds that:*

$$1-a(\varepsilon)\leq\frac{\frac{d^2Lg_n}{dw^2}(w)}{\frac{d^2Lg_n}{dw^2}(m_n)}\leq 1+a(\varepsilon), \tag{26}$$

where $a(\varepsilon)>0$ *and tends to zero for* $\varepsilon\to 0$.

(c3) *For any* $\rho>0$

$$\int_{B_\rho(m_n)}p(w|n,y^n,y_0,n_0)\,dw\to 1,\ \textit{for } n\to\infty.$$

Let ϕ_n be equal to $(w_n - m_n)/\sigma_n$, with $w_n \sim p(w|n,y^n,y_0,n_0)$. Then, given (c1) and (c2), (c3) is a necessary and sufficient condition for ϕ_n to converge in distribution to ϕ, where $p(\phi) = \mathcal{N}(\phi;0,1)$. ■

The proof of this lemma can be found in [4, Sec. 5.1]. Essentially, Lemma 4 states that, for large n, (c1),(c2) together ensure that inside a small neighborhood of m_n the function $p(w|n,y^n,y_0,n_0)$ becomes highly peaked and behaves as a normal density. Condition (c3) ensures that the probability outside any neighborhood of m_n becomes negligible for $n \to \infty$. Under these conditions, w has an asymptotic posterior limit $\mathcal{N}(w;m_n,\sigma_n^2)$.

Theorem 3. *Assume conditions in Lemma 4 hold*[9] *and that the gambles $g \in \mathscr{G}$ defined in Theorem 1 are integrable w.r.t. $p(w|n,y^n,y_0,n_0)$. Then, for the set of priors $\mathscr{M}$ in Theorem 1, (A.1) and (A.2) are always satisfied, while (A.3) and (A.4) hold if and only if $\mathscr{Y} = \mathbb{R}$ and, thus, $\inf\mathscr{Y} = -\infty$ and $\sup\mathscr{Y} = \infty$.* ■

Proof: *Since coherence and $\mathscr{G}_0$-prior ignorance properties do not depend on the likelihood (for coherence this holds since $p(y^n|w)$ is separately coherent), the fact that (A.1) and (A.2) are still verified is a direct consequence of Theorem 1.*[10] *First we prove the necessity of the conditions of the theorem, by showing that in the case $\inf\mathscr{Y} \neq -\infty$ or $\sup\mathscr{Y} \neq \infty$, (A.3)–(A.4) do not hold. Consider a gamble $g \in \mathscr{G}$ and the posterior $p(w|n,y^n,y_0,n_0)$ obtained in correspondence of the prior $p(w|n_0,y_0)$, which is equal to*

$$p(w|n,y^n,y_0,n_0) = \frac{\int_{\mathscr{Y}^n} p(y^n|y'^n)p(y'^n|w)p(w|n_0,y_0)\,dy'^n dw}{\int_{\mathscr{W}}\int_{\mathscr{Y}^n} p(y^n|y'^n)p(y'^n|w)p(w|n_0,y_0)\,dy'^n dw} \tag{27}$$

and can be rewritten as:

$$\frac{\int_{\mathscr{Y}^n} p(y^n|y'^n)\frac{\prod_{j=1}^n k(y'_j)k(n_0,y_0)}{k(n_p,y'_p)}p(w|n_p,y'_p)\,dy'^n dw}{\int_{\mathscr{W}}\int_{\mathscr{Y}^n} p(y^n|y'^n)\frac{\prod_{j=1}^n k(y'_j)k(n_0,y_0)}{k(n_p,y'_p)}p(w|n_p,y'_p)\,dy'^n dw}. \tag{28}$$

by using the fact that $L(y'^n|w)p(w|n_0,y_0) = p(y'^n)p(w|n_p,y'_p)$, with[11]

$$p(y'^n) = p(y'^n|n_0,y_0) = \prod_{j=1}^n k(y'_j)\frac{k(n_0,y_0)}{k(n_p,y'_p)}, \tag{29}$$

where $n_p = n+n_0$ and $y'_p = (n_0y_0 + \sum_{i=1}^n y'_i)/(n+n_0)$.
Consider the case in which $\mathscr{Y} = [0,1]$ (i.e, $\mathscr{Y}_m = \{0,1\}$ or $\mathscr{Y}_m = [0,1]$). Because of Lemma 3, for $y_0 \to 0$ ($y_0 \to 1$) and $y'_1,\cdots,y'_n \neq 0$ ($y'_1,\cdots,y'_n \neq 1$), it holds that $k(n_0,y_0)/k(n_p,y'_p) \to 0$ and, thus, that $p(y'^n) \to 0$ apart from the case in which $y'_1 = \cdots = y'_n = 0$ ($y'_1 = \cdots = y'_n = 1$) where the ratio $k(n_0,y_0)/k(n_p,y'_p) > 0$. Therefore, for $y_0 \to 0$, $p(y'^n)$ concentrates on $y'_1,\cdots,y'_n = 0$. From Lemma 3, it also follows that $p(w|n_p,y'_p)$ concentrates on $\underline{w}^*$ such that $db(w)/dw|_{w=\underline{w}^*} = 0$ when $y'_p \to 0$. Thus, for any choice of $\varepsilon > 0$, by continuity arguments, it is possible to find a $\underline{y}_0 \in Int(\mathscr{Y})$ and $\delta > 0$ such that

$$\int_{B_\varepsilon(\underline{w}^*)} p(w|n,y^n,y_0,n_0)\,dw > 1-\varepsilon,$$

for any $0 < y_0 \leq \underline{y}_0$ and $n > \delta$.[12] *In other words, for $y_0 \to 0$, the posterior $p(w|n,y^n,y_0,n_0)$ concentrates on $\underline{w}^*$. Similarly, for $y_0 \to 1$, the posterior $p(w|n,y^n,y_0,n_0)$ concentrates on $\overline{w}^*$ such that $db(w)/dw|_{w=\overline{w}^*} = 1$. Under continuity conditions for $g \in \mathscr{G}$ in a neighborhood of $\underline{w}^*$ ($\overline{w}^*$), this implies that, for $y_0 \to 0$ ($y_0 \to 1$), the posterior expectation of g, i.e., $E[g|n,y^n,n_0,y_0]$, concentrates on $g(\underline{w}^*)$ (on $g(\overline{w}^*)$).*[13] *Hence, for the continuous function $g = db(w)/dw$, since $g(w) = y_0$ and, thus, $g(\underline{w}^*) = 0$ and $g(\overline{w}^*) = 1$, it follows that $\underline{E}[g|n,y^n,y_0,n_0] = 0 = \underline{E}[g]$ for $y_0 \to 0$ and $\overline{E}[g|n,y^n,y_0,n_0] = 1 = \overline{E}[g]$ for $y_0 \to 1$, i.e., prior and posterior lower and upper expectations coincide. It can thus be concluded that (A.3) does not hold (no learning from data) and, consequently also (A.4) does not hold (no convergence).*
Consider now the case $\mathscr{Y} = [0,+\infty)$ (or $\mathscr{Y} = (-\infty,0]$), then if $y_0 \to 0$ the ratio $k(n_0,y_0)/k(n_p,y'_p) \to 0$ apart from the case in which also $\hat{y}'_n = 0$, where $y'_p \to 0$ and $k(n_0,y_0)/k(n_p,y'_p) > 0$. Therefore, for the same arguments of the case $\mathscr{Y} = [0,1]$, it follows that for $g = db(w)/dw$, $E[g|n,y^n,n_0,y_0] = g(\underline{w}^*) = 0$. This means that $\underline{E}[g|n,y^n,n_0,y_0] = 0$, it does not matter the value of y^n. Therefore, we conclude that (A.4) does not hold. (A.3) holds for some gambles. For instance, for the gamble $g = db(w)/dw$, (A.3) holds, since the upper expectation differs from its prior value for any $n > 0$ (but the lower expectation is always zero). Hence, the validity of (A.3) depends on the choice of the set $\mathscr{G}$. In particular, if $\mathscr{G}$ includes a function g which gets its infimum and supremum for $y_0 \to 0$ and, respectively, $y_0 \to \lim_{n\to\infty}\hat{y}'_n \neq 0$, then for $n \to \infty$ the prior lower and upper expectations coincide respectively with the posterior lower and upper expectations and, thus, (A.3) does not hold.
Finally assume that $\inf\mathscr{Y} = -\infty$, $\sup\mathscr{Y} = \infty$ and, thus, $\mathscr{Y} = (-\infty,\infty)$. Consider the parameters $n_p = n+n_0$ and $y'_p = (n_0y_0 + \sum_{i=1}^n y'_i)/(n+n_0)$ of the posterior density $p(w|n_p,y'_p)$. Under the conditions of Theorem 1, i.e., $y_0 \in Int(\mathscr{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$, it results that y'_p is bounded as in (14). From this fact it follows that conditions (c1) holds for any $y_0 \in Int(\mathscr{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$ since $y'_p \to y'_n$ for $n \to \infty$. For (c2), by continuity arguments is always possible to find an ε in the definition of (c2), for which (26) is satisfied for any $y_0 \in Int(\mathscr{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$ and, thus, for any prior in $\mathscr{M}$. It is in fact sufficient to consider the largest δ for which (26) holds for any y'_p in (14). This upper δ must exist finite, otherwise Lemma 4 cannot hold. Same considerations hold for (c3). Thus, for any prior in $\mathscr{M}$ satisfying hypotheses of Theorem (1), asymptotic normality holds. Under continuity conditions for $g \in \mathscr{G}$ in a small neighborhood of m_n, this implies that also (A.4) and, consequently, (A.3)

[9] This means that the imperfect observation mechanism still allows asymptotic normality to hold for any prior $p(w|n_0,y_0)$ with fixed $0 < n_0 < \infty$ and $y_0 \in Int(\mathscr{Y})$.

[10] More precisely, from Theorem 1, it can be derived that the likelihood $p(y^n|w)$, the set of priors $\mathscr{M}$ in the exponential family and the corresponding set of posteriors are strongly coherent.

[11] Equation (29) can be derived form (9).

[12] In the case $w^* = -\infty$, $B_\varepsilon(\underline{w}^*)$ must be intended as the open interval, e.g., $(-\infty,\underline{w}-1/\rho)$ for some $\underline{w} \in \mathscr{W}$.

[13] This was also proven in [8, Ths. 11–12].

hold. This proves that $\inf \mathscr{Y} = -\infty$ *and* $\sup \mathscr{Y} = \infty$ *are necessary and sufficient conditions for (A.3)–(A.4).* ■

The theorem states that for a set of Gaussian priors near-ignorance and learning/convergence are compatible even in the case of imperfect observations while this is for instance not the case for a set of Beta priors. The main point is that for the latter, when $\hat{y}'_n = 0$, $y'_p = (n_0 y_0 + n\hat{y}'_n)/(n_0 + n)$ can be made as close as desired to the left boundary of $Int(\mathscr{Y})$ and, thus, from Lemma 3 the posterior $p(w|n'_p, y'_p)$ can be made as closer as desired to a Dirac delta. Thus, in the integration in (27) the only meaningful term is the one relative to the case $\hat{y}'_n = 0$ and, therefore, $p(w|n, \hat{y}_n, n_0, y_0 = 0) = p(w|n'_p, y'_p = 0)$. Conversely, in the Gaussian case, since $|n_0 y_0| < \infty$ it follows that $|y'_p| = |n_0 y_0 + n\hat{y}'_n|/(n_0 + n) = \infty$ only if $|\hat{y}'_n| \to \infty$, but this case must have probability zero otherwise Lemma 4 would not be satisfied. This ensures that $p(w|n, \hat{y}_n, n_0, y_0)$ converges in distribution to $\mathscr{N}(w; m_n, \sigma_n^2)$ for any value of n_0, y_0 in Theorem 1. To better understand the peculiarity of the Gaussian density, assume that $p(y_i|y'_i) = \mathscr{N}(y_i; y'_i, \sigma_r^2)$[14], $L(y'_i|x) = \mathscr{N}(y'_i; x, \sigma^2)$ and consider

$$p(y^n|x) = \int_{y'^n \in \mathscr{Y}^n} \prod_{i=1}^{n} \mathscr{N}(y_i; y'_i, \sigma_r^2)\mathscr{N}(y'_i; x, \sigma^2)\, dy'^n. \tag{30}$$

Since $\mathscr{N}(y_i; y'_i, \sigma_r^2)\mathscr{N}(y'_i; x, \sigma^2)$ is equal to

$$\mathscr{N}(y_i; x, \sigma^2 + \sigma_r^2)\mathscr{N}(y'_i; \sigma_s^2(y_i/\sigma^2 + x/\sigma_r^2), \sigma_s^2),$$

where $\sigma_s^2 = \sigma^2\sigma_r^2/(\sigma^2 + \sigma_r^2)$, (30) becomes $p(y^n|x) = \prod_{i=1}^{n} \mathscr{N}(y_i; x, \sigma^2 + \sigma_r^2)$. Therefore, we can see that in this case the effect of the imperfect observation mechanism is just that of increasing the variance of the measurement noise.

7 Conclusions

This paper has discussed the problem of learning and prior near-ignorance for sets of priors in the one-parameter exponential family. In particular, for conjugate likelihood-prior models in the one-parameter exponential family of distributions, we show that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors $\mathscr{M}$ which guarantees prior ignorance without producing vacuous inferences. This result is obtained following both a behavioural and a sensitivity analysis interpretation of prior near-ignorance. We have also discussed the incompatibility of learning and prior near-ignorance for sets of priors in the one-parameter exponential family of distributions in the case of imperfect observations. In particular, we have shown that learning and prior near-ignorance are compatible under an imperfect observation mechanism provided that the support of the priors in $\mathscr{M}$ is the whole real axis. Future work will address the following issues: extension of the model to the multivariate case; extension to more general family of densities.

[14] This satisfies the hypotheses of Theorem 2.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-121785/1, 200020-134759/1 and by the Hasler Foundation grant n. 10030.

References

[1] A. Benavoli and M. Zaffalon. A model of prior ignorance for inferences in the one-parameter exponential family. Available at http://www.idsia.ch/∼alessio/TR2011.pdf.

[2] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, New York, 1985.

[3] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *Int. Journal of Approximate Reasoning*, pages 123–150, 2005.

[4] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley & Sons, 1994.

[5] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[6] Serafin Moral. Imprecise probabilities for representing ignorance about a parameter. *International Journal of Approximate Reasoning*, In Press, Corrected Proof, 2010.

[7] L.R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, pages 1–23, 1991.

[8] A. Piatti, M. Zaffalon, F. Trojani, and M. Hutter. Limits of learning about a categorical latent variable under prior near-ignorance. *Int. Journal of Approximate Reasoning*, 50(4):597–611, 2009.

[9] E. Quaeghebeur. Learning from samples using coherent lower previsions. PhD thesis, Ghent University, 2009.

[10] E. Quaeghebeur and G. De Cooman. Imprecise probability models for inference in exponential families. In *Proc. of ISIPTA'05*, pages 287–296, 2005.

[11] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[12] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24(4):463–483, 1997.

[13] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.